Folk Psychology,

Parallel Processing,

The Intentional Stance,

Supervenience,

Multiple Realizability,

And Some Hard Problems.

Or, how I learned to stop worrying and love the brain.

Gabriel Chapman

Philosophy Department,
Washington and Lee University

Advisor: Paul A. Gregory

May 10th, 2006.

The objective of this paper is to suggest that folk psychology is the most deeply entrenched myth in the history of the human race. Yet, it appears that – considered as a theory of how the mind works – we have ample reason to believe that folk psychology is inadequate and radically misleading. Its success hardly exceeds the bounds of the behavior it was engineered to explain, or behavior that has been socialized to conform to its categories. Folk psychology seems to bear little to no hope of being integrated, or making any significant contact with, our more fundamental physical sciences. Yet there exist professedly 'physicalist' philosophers who nevertheless propound the inviolability of folk psychology and the irreducibility of theories of the mental to theories of the physical. I want to suggest that their position is inherently tenuous. Folk psychology is usually defended hand-in-hand with anti-reductive positions on the mind-brain relationship, which often appeal to qualia and consciousness as especially 'hard problems' not amenable to physicalist explication. Both folk psychology and the hard problem are defended, I suspect, because it is thought that a fully materialistic perspective would not adequately characterize the breadth and depth of mental life. A story about the brain would not 'do justice' to the flights of fancy of the human mind. What I want to suggest is that folk psychology itself disguises the real depth and complexity of the capabilities of the human mind. Hence, I welcome Paul Churchland's model of mental representation, a model that does not rely on propositional, sentential representation. Along the way, his model will be briefly explicated, and arguments in favor of folk psychology will be addressed. Ultimately, it is suggested that part of the reason why we believe the mind is not amenable to physical description is that we have not yet, in our own minds, primed that domain for reduction. We have construed the domain itself, and what reduction should be, in such a way that reduction (of mental events to physical events) does not seem plausible. This has caused us to ignore the fact that the best way to understand the mind may be to incorporate both a bottom-up (neurophysiology) and a top-down (abstract, functional characterization) approach. Too many philosophers have supposed that the study of mind can proceed apart from its physical instantiation.

Parallel Distributed Processing

Before discussing the simple recurrent network, let us outline the architecture of a slightly simpler set up, the three-layer feed forward network. This organization features three layers of nodes – let us call them sensory, hidden, and output layers. The nodes in the first layer are each connected all of nodes in the second layer. Similarly, the hidden layer nodes project connections to the output layer. When signals originate at the sensory layer, the nodes achieve activation levels. These activation levels may be characterized in terms of amplitude, or, to be more faithful to the biological brain, in terms of firing frequencies. When one node makes a connection with a node in the subsequent layer, it conveys its activation level to the next node. But this value is inhibited or amplified by a synaptic weight at the point of connection. Thus, the activation level is transformed into a new value when it reaches this new node. This node will also be receiving connections from other nodes in the previous layer. The activation levels (affected by the synaptic weights) from each connection will be summed in this node into a new single value. This value represents the activation level of the node.

Considering each level of the network by itself, the simultaneous activation levels of the nodes at that level represent a *vector*. If our network had four nodes at its sensory layer, its pattern of activation at this level might be [2, 0.5, 0.9, 1.3]. After the effect of the synaptic weights and summing, this vector might be transformed into [1.4, 0.2, 0.7, 1.6] at the hidden layer. This transformation works just as well even if the next layer has a different number of nodes; in this case our four-dimensional vector would be transformed into a vector with a different number of elements. Since every node works on its portion of the vector simultaneously, these transformations can occur very rapidly. The network is said to be a parallel distributed processor. Unlike a traditional computer, this system does not need to serially perform millions of simple operations. This is auspicious because vector coding is an enormously powerful way to represent high-dimensional and complex sensory information and the parallel network suggests a way to perform computations on these complex vectors with a swiftness that biological, evolving creatures would find useful.

Adding recurrent connections to a network is done by establishing connections between higher level nodes and hidden-layer nodes. Since 'later' levels of the network now influence processing at the lower layers, the network now features a form of short-term memory. This memory gradually decays as several vectors course upwards through the system and the 'echo-effects' of previous vectors diminish. Short-term memory lends networks 'attentional' characteristics, and even the ability to deploy different conceptual frameworks against similar incoming data-vectors. Such networks can continue to process vectors even in the absence of external stimulus; that is, even without vectors beginning at the peripheral, sensory layer. Stimulation from recurrent connections to hidden layers enables the system to continue to propagate and process vectors, entirely on its own. Networks like these can display steerable attention: they can be primed so that recurrent connections will tend the hidden layers to process incoming vectors so as to activate certain *prototypes*. Prototypes are vectors (as measured at a certain hidden layer) containing values of a certain magnitude. After a network has been 'trained up,' the incident vectors it processes will create hidden-layer (the layer of nodes between the input and output layers) activation level that is in some subregion of the possible activation-vector space (a hyper-dimensional space) of the network. These subregions will be centered on 'prototypes,' which are paradigmatic vectors that signify certain characteristics, like prototypical 'mine' or prototypical 'rock' vectors. According to the connectionist way of speaking, the 'concepts' and 'theories' that a network brings to bear on a situation are described by the way in which its possible hidden-unit activation vector space is 'divided up:' and it is divided because the prototypes which center these subregions act as 'sinks,' to some extent pulling incident vectors towards them.[1] The divisions and prototypes – like 'theories' and 'concepts' thus have an impact on what is representable by the network. The subregions and prototypes shape the vectors as they course up through the network.

Moreover, recurrent networks have genuinely dynamical properties. Jeffrey Elman, in "Language as Dynamical System," shows a very simply recurrent network can perform the relatively complex task of predicting the subsequent word in a sentence from the preceding words which lead up to it. Importantly, the network configures its

[1] Later discussion of Hebbian learning mechanisms and the 'ampliative' effects of information processing will make clearer why this is so.

activation space (the vectors consisting of a certain number of dimensions) in a way that represents the grammatical properties of the words that it works with. The network has inferred conceptual similarity and difference among the words. More extensive treatment of Elman's work will have to wait for a later date; but one lesson which should be taken away from this is that the words that fed into the network are better understood as *operators* rather than *operands*. Rather than being represented and 'computed on' by the system, the words actually drive the processing in a more direct fashion. From the dynamic perspective, this is an important distinction to make.

Consider, for instance, color recognition. The retina features three types of cones, each of which is sensitive to a different range of wavelengths. A color can then be coded by a unique three-element activation pattern. Vector coding and parallel processing can scale up to far higher dimensions with no loss of computational speed: witness: motor coordination and *propioception*, the means by which the brain senses the limb's orientation in space. A vector representing the contraction state of all the body's muscles would have thousands of dimensions. The cerebellum, known to be instrumental in muscle coordination, appears to have the sort of massively parallel architecture described above.

Now, if the output vectors can be systematically related to the input vectors, parallel networks will be able to perform biologically and functionally useful operations. At first, the outputs of an artificial network are random, and bear no systematic relation to the inputs. The solution is to adjust the synaptic weights so that the desired outputs for given inputs are achieved. The traditional technique for adjusting the weights is called back propagation of error, or the generalized delta rule. The output achieved is compared to the desired output, and changes are made to the synaptic weights in each layer. Unfortunately, this learning mechanism requires a 'teacher' who can recognize the difference between actual and desired output, and thence tweak dozens or hundreds of synaptic weights to push the output closer to the desired value. This is easy do when a conventional computer program is simulating a parallel network, but obviously evolving creatures do not have the benefit of a teacher to calculate the difference between the actual and desired values. Back propagation methods also suffer from increasing learning time as the network gets bigger and the changes must be distributed. Luckily, there are alternative learning techniques that adjust the synaptic weights in a more local fashion, and do not rely on a teacher that knows the desired output value. In Hebbian learning, a synaptic weight is increased (the multiplier increases in value) if a high activation level at an earlier node is coincident with a high activation level in the post-synaptic node (Churchland 1989, pp. 246-247).

Yet this type of weight-adjustment seems a far less sure thing. Why would the coincidence of high levels of activation at pre- and post-synaptic nodes indicate that that synapse ought to be increased in weight? It seems that this simultaneous activation could simply be a coincidence, and increasing the synaptic weight in response to it would not necessarily bring the output vector closer to the desired value. At this individual synaptic level, this may well be true, but keep in mind that the downstream node is also contacted by many other presynaptic nodes, which are in turn being activated by recurrent nodes. Therefore a coincidence of high levels of pre- and post-synaptic activation is likely to indicate that the overall level of activation in that nodes' neighborhood is high, and this is not likely to be a 'chance' occurrence. This part of the network is probably hitting upon a salient feature of the vectors that are incident upon it. Hebbian learning appears to encourage and increase activation in those areas which are being highly stimulated; and these are just the areas likely to be representing an important and recurring component in the stimuli (incident vectors). An important feature of parallel networks is that they can amplify important features of a noisy and confused stimulus (think about hearing a loved one's voice calling out your name at a noisy football game) and Hebbian learning provides an explanation for how networks adjust themselves so as to 'amplify' important elements of a stimulus. Moreover, even though Hebbian learning may not optimize every synapse as readily as back-propogation techniques do, any learning technique need only optimize patterns of activation across the network in order to be effective. The destruction of individual nodes in the network, or the improper adjustment of their weights, will have only a very minor effect on output. Like real brains, parallel networks are relatively damage and fault tolerant, especially in comparison to digital computers.

Although the weight-adjustment mechanisms in real brains remain obscure, the parallel processing model appears to be vindicated in the neuronal structure of the cerebellum, an evolutionarily primordial part of the brain. Here is a system that is instrumental in regulating the complex coordination of our muscle systems. Parallel processing systems can rapidly perform transformations on data that can encode enormously complex, many thousand-dimensioned information. Compare the difficulty conventional artificial intelligence has had with developing bipedal walking robots with the lithe and graceful movements of the evolved creatures all around us. These programs have been frustrated with being able to compute and react fast enough, even though silicon-based computers propagate their signals many millions of times faster than the biological nodes of brains, which require around a tenth of a second to do so. This suggests that the computational strategy being used in current artificial intelligence programs is the wrong one.

In other ways, the aptitudes of parallel networks seem to mimic our own. They are quick and effective in performing complex recognition tasks, like discerning 3-D shape from 2-D shading, or using sonar echoes to discriminate mines from rocks (Churchland 1989, p. 164). These types of perception tasks are often just those that frustrate traditional serial computers, or require an inordinate amount of processing time. But just as importantly, recurrent networks have trouble with the same sort of tasks that real human beings do. Like us, they are less adept at recursive computations, the types of tasks that serial computers are best suited for. That is, neither real brains nor recurrent networks manipulate symbols according to instructions, as conventional computers are programmed to do. The prototype-activation model suggests that what we call explanatory understanding is really an outgrowth of faculties originally developed for perceptual recognition. If we are interested in maintaining continuity with our evolutionary history, this seems to be an attractive position. All of our capabilities must, at some level, be smooth outgrowths brain structures that developed in response to the demands on evolving creatures to flee, fight, reproduce, and forage. There must be an evolutionarily plausible etiological story for even our high-level capabilities, like those we call 'inference' and 'explanatory understanding.' An evolutionary, etiological story for these capacities has been largely withheld on the grounds that such linguistically-dependent activities are on one side of a great gulf that exists been linguistic and non-linguistic creatures.

Dynamics

Dynamics allows us to gain descriptive traction on systems of very high dimensionality; systems that from a finer-grained perspective have an enormous number of variables influencing their behavior. The mind-system certainly seems liable to an overabundance of variables.

The most cursory examination of just one player in the mind-system, the brain, will reveal the magnitude of the task if we adopt too fine-grained a view. Containing approximately 10^8 neurons (not counting glial cells), and each enjoying about 10^3 synaptic connections, the brain houses over 10^{11} synaptic connections. Each one of these represents a variable with a distinct weight value, though many of them may be coordinated with one another, such that their values are not wholly independently determined. The sheer volume of information incident from the sensory pathways, and the extent of communication between the cerebral hemispheres only further dramatize the issue. Often, it seems that opposition to the physical stance is partially motivated by the fear that this type of description's complexity would spiral out of control. The question remains whether it is possible to adopt exclusively physical description and still have a level of description that would be tenable for human beings to grasp.

Are our minds are too complex to (really) understand themselves?

Dynamical description remains relatively undaunted by complexity of high magnitudes because it deliberately aims to simplify: "...conceptually understandable models are sure to be greatly simplified in comparison with real systems. The goal is then to look for simplified models that are nevertheless useful." (Port, 1995, p. 47) The question arises: what can we hope to gain from dynamics and connectionism, as opposed to other ways of understanding the human mind? We would hope, with a new system of understanding, to address the limitations of our current folk-psychology, which has made little headway in: 1) learning, 2) the cognition of non-linguistic creatures, or cognition that does not involve linguistic representation, and 3) non-prototypical human cases including children and people with cognitive deficits or brain lesions. The principle desirable characteristic of folk psychology is that it makes it possible for us to discuss and predict (very narrowly) people's behavior through the medium of our public language. Folk psychology distills the situation in such a way that we can, through the shorthand of folk psychology and public language, discuss people's behavior. We can generalize over what over what is undoubtedly a greatly more complex actual situation: a hugely complex physical system (the brain) interacting with a complex physical environment. Defenders of the folk psychological framework often question how it would be possible to talk about mental systems in any other way. Dale Jacquette daunts us with the following example:

> What if a history of the Watergate scandal were to be given in a book filled with nothing but chemical formulas describing the brain and other physical events that took place at the time involving participants in the break-in, wire-tapping, and cover-up? The test would be to have the most knowledgeable scientists who have not been informed about the psychological and social correlates of behavioral-materialist-functional elimination or reduction examine these formulas and give an interpretation

> of the events described…the property dualist and folk psychologist are in a better position to explain these events by appealing to agents' thoughts and intentions. (Crumley, 2000, p. 41-2)

Now, I grant that folk psychology offers a more facile way for us to describe the Watergate scandal and the behavior and internal states of the players involved. But my intuition is that folk psychology wins us ease at the expense of accuracy. We should not remain complacent in an existing conceptual framework simply because we fear that the alternative is too complex.

Suppose that dynamics were to offer up a model which would enable us to describe the Watergate situation, including the activity of its participants. Would this description avoid dealing specifically with the internal states of the participants? Perhaps internal states would simply be subsumed under the more general dynamic description of the environment and the participants. If internal states were thus concealed, we would consider this a disadvantage of the new description. Behaviorism was upbraided for refusing to confront internal states head-on, for refusing even to discuss their existence. But since it was plain for all to see that we have subjective internal states, a theory's refusing to acknowledge these states is usually considered a reductio of that position. Surely, it would be unfair to ask a new dynamical description to ascribe 'beliefs', 'desires', 'fears', and 'objectives,' but we would like it to say something *distinct* about internal states. However, one might question why this demand for distinct treatment of internal states is warranted. Part of folk psychology's grip on our mind-theorizing seems to lie in this insistence.

The insistence that particular attention be paid to internal states seems to grow out of a generally Cartesian predisposition. As Timothy van Gelder points out, "In the Cartesian framework, the basic relation of mind to the world is one of representing it and thinking about it, with occasional 'peripheral' interaction via perceiving and acting." (Haugeland, 446) On this account, it is entirely possible to imagine the internal system existing independently of the external environment, although there are 'gates' through which the two can interact: namely the sensory transducers and motor system. The internal system's interaction with the external environment is thus discrete and piecemeal. The brain could be disconnected from the external world and would remain fully a mental system, conducting operations on representations. The brain in the vat, from this Cartesian perspective, fully counts as a mental system.

On the other hand, the dynamic perspective holds that insulating the brain from the external environment destroys the mental system of which both are a part. The brain constitutes the portion of the mind that is causally efficacious in our behavior; but by itself it cannot be a mind. Van Gelder thus uses 'cognition' to refer to the brain's specific role in our behavior (Haugeland, 447); thus cognition comprises only a portion of what we describe as 'mental' or 'mental systems.' The internal system is irrevocably *coupled* with the environment, since its variables interact causally and reciprocally with external variables. The brain only becomes part of a mind-system when it is environmentally situated.

Folk psychology elaborates extensively on the internal states of the participants, in order to make the more observable features of the situation (including the participant's overt behavior) more intelligible. From within folk psychology's framework, we can develop a relatively coherent narrative. We can tell a story about what went on during the Watergate days. We should, however, be wary of the apparent plausibility and consistency of this folk-psychological story, since we are so accustomed to stories of this nature. It seems that Jacquette sets up a false dilemma: either you can have this neat and tidy folk psychological story, or you must somehow cope with interminable equations. And, he adds, unless you are going to relate these physical events back to their folk-psychological correlates, it will be impossible for you to 'interpret the events.'

The interpretation Jacquette has in mind will require the resources of public language. This is how we would verify the scientists' understanding. But public language has been engineered in light of folk psychology. The scientist may well look at the physical event description and realize for himself what happened, but he would be unable to describe via the language-medium what has occurred. The reason the scientists' successful interpretation seems unlikely is because we are demanding that he interpret and report using a public language riddled with folk psychology. Jacquette's expectations regarding the scientist's failure speak only to the scientist's possible inability to verbalize his folk psychology-free understanding.

Another element at work in Jacquette's example is his attempt to daunt us with the sheer magnitude of the task facing the scientist: somehow she must evaluate a huge list of chemical equations and organize them, in her mind, in a coherent way. Jacquette expects our scientist to be overwhelmed by micro-level description. Dynamics, however, suggests an alternative method of description that can avoid this problem. It may be able to avoid becoming bogged-down in micro-level description.

Our search for a new means of description is motivated by a fear that folk psychology is radically misleading in way it represents how people work. The question is whether folk-psychological narratives pick out objective features of the phenomena and actors being described. Or, are these narratives merely economical means by which human beings lay a shroud over the real goings-on, a tacit and unwitting agreement made among all language-users, so that oneself and others are easier to talk about?

Dennett's Intentional Stance

Daniel Dennett holds that we can validly assume the intentional stance towards a thing if its behavior is 'reliably and voluminously' predictable if we ascribe it intentional states (like beliefs and desires) and hypothesize that it will act in light of them. The intentional stance is an alternative to the physical stance, where one, "determines its physical constitution (perhaps all the way down to the microphysical level) and the physical nature of the impingements upon it, and use your knowledge of the law of physics to predict the outcome for any input" (Crumley, p. 228). Dennett's thought experiments involving super-physicist aliens seem to suggest that he considers the physical stance's most accurate iteration to lie at the microphysical level. Indeed, this may be the most accurate form of physical description, but perhaps a compromise can be made between accuracy and utility such that the physical description achieves both good accuracy and practicality of use. Like Jacquette, Dennett does not seem to consider the possibility of attending to more global regularities in a way apart from that apprehended in folk psychology. He fails to suggest any alternative strategies that might hit upon regularities that better represent the true goings-on.

Dennett's 'intentional stance' may seem to trample on the 'staunch realist' intuition that whatever is actually true of a thing is logically independent of whatever we find useful to say about it, for the ease of our own prediction and description (Clark, 2001, p. 56). But Dennett stumps for the intentional stance by seeking to demonstrate its power, and even, its unavoidability. Beyond this, Dennett believes the intentional stance can gain real, objective ground by discovering real patterns and features in the environment that are not dependent on any individual observer. Consider now two thought-experiments comparing an alien super-physicist, who has taken the physical stance to its most accurate, micro-level description, and a humble human, armed only with a normal folk psychological understanding.

> Suppose, that is, that they did not need the intentional stance – or even the design stance – to predict our behavior in all its detail. They can be supposed to be Laplacean super-physicists, capable of comprehending the activity on Wall Street, for instance, at the microphysical level…But if the Martians do not see that indefinitely many different patterns of finger motions and vocal cord vibrations – even the motions of definitely many individuals – could have been substituted for the actual particulars without perturbing the subsequent operation of the market, then they have failed to see a real pattern in the world they are observing. (Crumley, p. 232-3)

In the next thought-experiment, a Martian super-physicist and a normal adult human are both observing a scene. A homemaker receives a phone call from her husband telling her that he will be bringing the boss home to dinner. The wife tells him to pick up a bottle of wine and drive safely. The human is able to predict that, barring unusual circumstances, a vehicle will arrive at the house with two people in it, one of which will be carrying a bag with a bottle containing an alcoholic beverage. The wife will have a dinner just about ready to be set out (Crumley, 235).

> The Martian makes the same prediction, but has to avail himself of much more information about an extraordinary number of interactions of which, so far as he can tell, the Earthling is entirely ignorant…The Earthling's performance would look like magic! How did the Earthling know that the human being who got out of the car and got the bottle in the wine shop would get back in? The coming true of the Earthling's prediction, after all the vagaries, intersections, and branches in the paths charted by the Martian, would seem to anyone bereft of the intentional strategy as marvelous and inexplicable…There are patterns in human affairs that impose themselves, not quite inexorably but with great vigor, absorbing physical perturbations and variations that might as well be considered random; these are the patterns that we characterize in terms of the beliefs, desires, and intentions of rational agents. (233)

Dennett claims that the human armed with folk psychology would be able to see a real, objective pattern in the events unfolding that the Martian may have missed. The Martian will have failed to see the pattern if it is incapable of distinguishing the relative importance of the physical micro- and macro- events it is observing. The human recognizes both episodes – the buy-order for 500 shares of stock, and the arrival at the house with wine – could have unfolded in myriad different ways. The Martian would recognize these different cases as having a great degree of concordance in particulars, but he might fail to recognize that, at a certain level of description, these are versions of the *same* event. Suppose that in version 1 of the dinner episode, the husband takes his usual route home. In version 2 he goes an alternate route – suppose the normal route is closed due to a traffic accident. The alien might fail to recognize that the driver's taking a different way home has no significance, with respect to the way his actions are being informed by his disposition (I hesitate to say 'objectives', 'beliefs', 'desires', etc.). This change would cause massive microphysical differences, and therefore lead the alien to suppose that two versions of this episode are substantially different. The alien will have missed a pattern in the two episodes.

Moreover, the super-physicist alien might be unable to predict the consequences of counterfactual, unusual, or imagined events that we could easily and quickly detail. Suppose that the Secretary of State were to announce he was a paid operative of the KGB (Crumley, 232). Although an unprecedented event, we would be able to predict press conferences, splashy news headlines, investigations, and resignations. "Note," says Dennett, "…it describes an arc of causation in space-time that could not be predicted under any description by any imaginable practical extension of physics or biology." (232)

I fear, however, that there is something unfairly aiding and abetting the human's predictive performance. Our human observer, after all, is observing *humans* going about their business. Perhaps their behavior is itself informed by their acquaintance with folk psychology: their beliefs that they have beliefs, desires, fears, hopes, and the like. These ascriptions may be learned in the personal, subjective case, just as much as they seem to be in our talent for ascribing states to other people. As we mature, we become increasingly astute in attributing states to people, and this in turn increases our predictive and explanatory prowess. What began as a primitive distinction between happy and sad/angry matures into the ability to recognize shades of gray and mixed emotions. We learn to distinguish between consternation and anger and become more sensitive to contextual issues. Likewise we seem to become more fine-grained in our own cases. Perhaps the intentional stance is a self-fulfilling prophecy: it works as well as it does just because it deals with agents who inform their behavior using the same system. This seems to give the intentional stance the objective grip that it needs, but oddly so because the whole matter hinges upon a contingent way in which humans understand themselves and others. If the intentional stance alights upon patterns in human behavior, the intentional stance describes something more akin to a *convention* than a physical regularity.

Supervenience

One early variant of physical reductionism that has since fallen out of favor is type-identity theory. Type-identity theorists held that a mental kind is identical to a specific neurobiological configuration. However, this view was roundly criticized for 'species chauvinism,' because it claimed that in order to have pain (or more sophisticated mental contents like beliefs, desires, and the like) amounted to having the brain state that characterizes human beings in such a mental state. But clearly, other species of animals with radically different neurophysiologies (like octopi) can experience pain, and perhaps other terrestrial animals can experience some items that feature in the belief-desire psychology. Moreover, we cannot exclude the possibility that extraterrestrial creatures with completely alien physiologies might enjoy equally rich mental lives: might have beliefs, desires, and the like. The point may be made even within a single species. Given widely different experiences and upbringings, and the possibility of abnormalities or lesions that do not lead to substantial cognitive deficit, it seems that very similar mental states can be achieved in human beings that feature fairly substantial differences in their neural networks. So it seems absurd to propose that a mental kind can be identified with a specific brain state within a single species, and *a fortiori* across species.

Rejection of type-identity theory does not entail rejection of token-identity. This position holds that for any given mental event (at time T), there is a unique physical state of affairs (in human beings, a neurophysiological state) that is the instantiation of this specific instance of a mental kind. Commitment to token identity is entailed by a rejection of dualism, in that it rejects the notion that there is anything above and beyond this particular physical instantiation at time T. Talk of mental kinds is meant to generalize beyond particular instances, but a complete description of a particular instance can be achieved by a unique description of the physical state of affairs at that time.

I think there ought to be concern about whether it is wise to attempt to talk about mental kinds across species. That is, one may feel the attraction of talking about pain (for octopi) or pain (for humans) rather than pain simpliciter (which can be differently instantiated). To what extent can various species be said to be enacting their version of the same state? Any non-species specific mental kind would have to incorporate only those behavioral, functional, and phenomenological aspects which are common to all specific instances of the kind. One might suspect that, in fact, there are no commonalities at all; certainly with respect to phenomenological properties, but also very likely with respect to functional properties. The nature of motor output will vary not just with respect to the differences in the central nervous system, but also due to variances in physical conformation and locomotion and communication systems.

To achieve a non-species specific mental kind, we may have to keep our description of it fairly basic and coarse grained. For instance, by watching the behavior of the silicon alien go about its business (reversing, Dennett's example), I may come to the conclusion that it has beliefs about the location of objects and the motivations and intentions of the aliens around it. I may even be able to guess at the content of this alien's beliefs and intentions, but I infer all of this only from a very coarse-grained view of the alien's behavioral output.

Doubt about the possibility of non-species specific mental kinds is somewhat ill-founded, as suggested by the following consideration: as was mentioned earlier, we should expect that there is a good deal of neurophysiological difference among the members of our own species. Yet no one is tempted to become a mental-kind solipsist, even if no ever instantiates exactly the same global brain configuration as anyone else. Consider some of the more extreme cases: those who have had their corpus collosum severed (such that their cerebral hemispheres cannot communicate directly with one another, but actually cue one another using the motor system, as if communicating with a different person), or those who lack (or have lost) one of their cerebral hemispheres. After observing subjects like these perform tests and interact with others, we do not doubt that they enjoy the full menu of common mental kinds; *even though they never instantiate the same (or even approximately the same) global*[2] *neurophysiological configuration as I do.* Concern about the possibility of general mental kinds seems to come in degrees; and judgment about a mental kinds' applicability seems more dependent on behavior than physical brain state. I am probably never tempted to grant a mollusk (or even a comatose person) any of these mental kinds, I might do so after interacting with a robot or a silicon alien. I may be tempted to in spite of the fact that my physical provenance is far closer to the mollusk and comatose person than the alien or android. Abstract-functional organization and behavioral output seem far more important than evolutionary-biological extraction when it comes to assigning mental kinds. The behavior of creatures very differently physically organized may be similar enough to warrant us in assigning them roughly similar mental kinds – ascriptions that will help us explain and predict their behavior. Moreover, doing so will enable us to predict the behavior of psychological creatures without a need to pay attention to their specific physical makeup: "If any creature is in this situation, and it has the belief that Q, then it will be likely to do Y." Maintaining mental kinds has the putative advantage of generalizability and simplicity.

Some philosophers, therefore, wish to reject type-identity, on the basis of the plausibility of multiple-realization, while at the same time maintaining token-identity. They are committed physicalists in the sense that they believe fully physical descriptions (without appeal to mental kinds) are complete descriptions of individual, datable events; even if such a description is only in principle possible. Yet they hold fast to mental kinds on the grounds that they offer generalized descriptions (and therefore are explanatorily and epistemologically useful) that cannot be captured in purely physical description, even if the physical stuff is the only 'real' stuff in the universe.

My intuition is that appeal to mental kinds may only achieve generalizability in cases where it is doubtful that the same mental kind is really involved: e.g., in pain (for a human) versus pain (for an octopus). These mental kinds may be similar only for the minimal reason that they describe roughly equivalent behavorial input-output relations. Moreover, I believe there is strong reason to think that kinds like 'pain' represent the best case scenario for inter-species and inter-subjective mental kinds. Perhaps we have greater reason to suspect that more content-oriented and dispositional kinds (e.g. beliefs, desires, fears) are less likely to be subsumed under similar descriptions (mental kinds) across species, and across subgroups (e.g. cultural or age) of a given species. Moreover, maintaining that one is a committed physicalist seems to be at odds with claiming that one expects mental kinds to be preserved no matter how neuroscience develops in the future seems to be at odds with claiming that one is a committed physicalist. If mental kinds are insulated and immune to revision, then one seems to claim that they form a distinct category that is not dependent on developments of understanding about the natural world, and that mental kinds therefore form a distinct and independent ontological category beyond the 'physical stuff.' Perhaps this concern is misguided, and to help

[2] It may be suggested that what is relevant is not one's global neurophysiological configuration, but rather the configuration of specialized sub-areas of the brain. These may function essentially normally in either of the patients described, so that we would not expect to see differences between them and the 'normal' patient. However, if we take global coordination to be relevant to the synthesis of consciousness, and if we note the importance of subcortical brain structures (which often project to disparate areas of the brain), it seems that anything which radically affects global communication must thereby affect the operation of the sub-areas as well.

explicate the relationship between mental kinds and the physical stuff, we should look at the concept of supervenience.

Minimally, and most generally, supervenience describes the relationship that holds when a world alike in all A-respects must therefore be alike in all B-respects. A-properties are said to *supervene* on B-properties if and only if sameness in B-respects guarantees sameness in A-respects. A simpler formulation (suggested by Jaegwon Kim) essentially does the same work: worlds alike in B-facts must also be alike in A-facts (Kim, 2003, p. 562). However, when the subvenient set of facts (B-facts) is 'all the physical facts (in the world),' this global supervenience relation takes on the character of a vacuous truth. Mental properties globally supervene on physical properties if and only if worlds alike in all physical respects must therefore be alike in all mental respects. But of course this is true, since worlds alike in all physical respects must be one and the same world.

The concept of supervenience is also intended to capture an asymmetrical dependency between the sets of facts: the supervenient set of facts is supposed to be dependent on the subvenient, but not vice-versa. Global supervenience can do this, because it seems to imply that there cannot be a change in the supervenient facts without *some* change in the subvenient. But since we are dealing with the relation between mental facts and the *whole set* of physical facts, we need to elaborate, lest the concept be limited to the vacuous point that the same world will have all the same properties and facts. The most natural point of elaboration seems to be located at the asymmetrical dependency: what types of changes in the subvenient facts correspond with changes in the supervenient facts?

As we have seen, global supervenience is holistic (Kim, 563), in that it does not make claims about specific mental kinds and their relation to specific physical configurations. In this sense it is more of a minimal metaphysical commitment. Another variety of supervenience, which Jaegwon Kim calls 'strong supervenience,' describes the relationship between individual super- and subvenient properties, "…such that every supervenient property has at least one base property that is sufficient for it…" (Kim, 563). However, if strong supervenience is true of the two sets of facts, it seems that the supervenient set is in danger of being reduced to the subvenient. Thus non-reductive physicalists will be hard pressed to make a claim more interesting than global supervenience without thereby raising the specter of reduction. Non-reductive physicalists, in appealing to supervenience, would like to commit to more than simple covariance. Covariance could entail that each set of facts is both super- and sub-venient on the other set; but they want to describe an asymmetrical dependency. Is there a way to describe an asymmetrical dependency that does not lead us part of the way to reductionism?

Multiple-Realizability

Nonreductive physicalists might hold, however, that the specter of reductionism is never to be feared. Although we may worry that supervenience (or any other possible mental-physical relation) threatens reductionism, type-reduction will never, and can never, be accomplished because it is impossible for us to know all of the possible physical instantiations of a given mental kind. For we would have to discover and list the instantiations that currently (or have ever) existed in the universe, and those which are possible but will never be. Since we can never know all of the possible physical instantiations, it would be impossible for us to come up with the disjunctive physical predicate which would be necessary and sufficient for a given mental kind. That is, it is impossible for us to come up with the biconditional bridge laws which are supposed to be necessary for reduction of these mental kinds.

Moreover, some (see Marras, 1993, p. 287) claim that supervenience is, at its core, antithetical to type-physicalism. Marras says that the asymmetrical dependence captured by supervenience is a sort of ontological dependence. But ontological dependence must necessarily entail that the two kinds are distinct, for one to be dependent on the other. If supervenience limited itself to simple covariation, the possibility of the properties (mental and physical) being identical would remain; but it is precisely the asymmetrical dependency claim which enables Marras to say that supervenience must, at its core, hold that the two phenomena are distinct.

Jaegwon Kim holds two main criticisms against the foregoing account of supervenience: first, it relies on an outmoded model of reduction (specifically, one that demands biconditional bridge laws), and second, that supervenience leads to the apparent over determination of mental events, implying that one of the putative causes (namely, the mental one) may be eliminated.

The nonreductive physicalists hold that the required biconditional bridge laws are not in the offing because (for them to hold across nomologically possible worlds) the physical predicates would have to be very long disjunctive statements, where each disjunct represents a sufficient instantiation of the mental kind at issue. These disjunctive statements are probably forever closed to our discovery, since in the first place we could never know if they were complete. Supposing we were able to complete them, they would be so unwieldy as to be of no explanatory use. As Marras puts it, “there can be no explanation or reduction without conceptual or representational economy.” (284) Since the ‘bridge’ between the mental and physical kinds are laws, they are supposed to hold within any possible world that abides by the ‘basic’ laws (i.e. laws of physics) of this world. Thus the anti-reductivist can demand an impossible task of he who seeks to reduce the mental property: come up with a disjunctive predicate which contains *all* of the possible physical instantiations of this mental kind (such that the law will hold across all nomologically possible worlds). Then he points out that this predicate can be of no use in a reasonably economical explanatory and predicative framework, and, hence, the mental kinds ought to be preserved for at least these reasons. Granted, these ‘fat’ P*properties (the disjunctive predicates) would be nomologically coextensive with the mental kinds they are biconditionally related to, but for Marras this is only a necessary (but not sufficient) condition for reduction of the property. The additional conditions, according to him, are epistemic: “…explanatory potential, predictive use, degree of confirmation, deductive integration, and the like.” (282)

To this point, I have understood reducibility to be primarily an ontological issue: it is supposed to be a point about what there ‘really’ is. The aspiring reducer of mental kinds might acknowledge their usefulness as a heuristic or investigative tool, but maintain that they can be nothing but shorthand for the ‘real’ underlying physical activity. In sense, all materialists (including non-reductive ones) are committed to this claim, since they grant that token-reducibility is possible – in effect, that physical stuff is all that there is. On this interpretation, it seems illegitimate for Marras to introduce epistemic conditions. In the first place, these epistemic conditions seem to be relative to the idiosyncrasies of the human mind, and it does not seem right to admit these issues into a question of whether one kind of event "really just is" another sort of event. Aside from the obvious imprecision in the terms at use here, it seems clear that this account of reducibility is inadequate. Appeal to the 'ultimate stuff' or the 'stuff there really is' seems naïve. Reducibility is also a question of whether one type of event can be explained in the terms of another event. So Marras may legitimately introduce considerations that reflect the idiosyncrasies and limits of the human mind. We have to keep in mind, as well, that non-reductionists like Marras acknowledge that a given mental event (e.g. at time T) could be wholly explicated in physical terms. The sticking point for non-reductionists is that the mental *kind* instantiated at time T could not be wholly explicated by a very large disjunction of different physical instantiations. This could not be achieved in a way that would be useful for our explanatory purposes; this disjunction could not 'fill in' for the mental term. This is why Marras says that epistemic conditions may be admitted as conditions on reducibility. This is why one kind cannot be said to have 'replaced' or 'reduced' another kind until it has been shown to meet the same epistemic requirements.

Marras is warranted in imposing epistemic conditions on reducibility, but he may be less justified in relying on the deductive-nomological model of reduction. I am suspicious of the whole structure of the 'multiple instantiation' rejection of reducibility of mental kinds. Instead of excoriating neurological research, the anti-reductionists make a point about the great variety of *imaginable* physical instantiations of sophisticated psychological activity. They then rely on the deductive-nomological model of reduction to point out that it would be impossible to list all of the possible physical instantiations, such that the disjunction of these physical realizers would be nomically coextensive with the mental kind they realize. Aside from the practical matter of assembly the disjuncts, disjunctive laws often come for criticism, thus further reducing the desirability of a deductive-nomological reduction. Since the biconditional bridge laws are not in the offing, it appears impossible to derive the mental kinds from the disjunctive physical predicates, using the bridge laws as auxiliary premises.

But this is not the way one expects the rejection of reducibility to proceed. This argument seems suspiciously *a priori*, just when one expects the verdict to be empirically based. One may think that multiple-realizability does most of the work here, but in fact the deductive-nomological model does the majority of the work. In particular, its demand for nomically coextensive predicates seems to make the anti-reductive decision a foregone conclusion.

The first premise concerning the possibility of multiple realization is largely unassailable. Few would want to claim that no terrestrial creatures aside from human beings instantiate mental kinds, and fewer still would disallow the possibility that there exists intelligent life elsewhere in the universe. However, I would suggest that the point about the multiple-realizability of mental kinds is questionable for the following reason. What cause have we to suppose that the mental kinds 'pain,' 'belief that p,' and 'fear of X,' can in fact be multiply instantiated? That is, how do we know that the *same* mental kind is being differently instantiated? Why are not these mental kinds different for each species, whether silicon-based alien, human being, or any other terrestrial creature? Why are not we talking about 'pain$_{\text{for humans}}$' being instantiated in the lateral cortex, while 'pain$_{\text{for aliens}}$' is realized in the region A12 of their silicon central nervous system? Why do we suppose that a meta-kind of 'pain' (without a subscript designator) can include these different kinds of pain? The assumption is that the meta-mental kinds (those not particular to any species) capture those factors which are common to each instantiation.

When the argument against reducibility invokes multiple-realizability, it relies upon the idea that mental kinds can be attributed across species and even nonbiological physical bases. Thus, when we speak of two different creatures or systems having 'pain' or 'belief that p,' we are must be talking, to a certain extent, about the same kind of thing. Doing so seems compatible with granting that that there might be differences in features of each instantiation; we need only suppose that there are enough important common features. Since the term for the 'meta' mental kind will be shorthand for these features, we do no know what we talking about until we have discovered them. Agitating against reductionism, Ausonio Marras points out that what we want in a reduction are not long disjunctive predicates, but, "…'perspicuous' predicates that effectively represent the common, unifying features of the objects in their range, in terms of which we can then explain their common power of, say, determining a given mental property (for each mental property). There can be no explanation or reduction without conceptual or representational economy." (284) Granted; but Marras' argument presupposes that the common features of mental properties have already been unveiled. Is this so?

What are the common features which identify 'meta' mental kinds? A good place to start seems to be their functional descriptions. Insofar as we can talk about pain across species, we are probably talking about common behavioral traits that suggest possession of this mental state; like wincing-type expressions/yelping and avoidance behavior. However, there are problems in trying to characterize the functional/behavioral characteristics of a given mental kind. On the one hand, we do not want to be too specific about the functional relationships. Clearly, in order to achieve the cross-system applicability, we cannot describe a meta-mental kind as causing a specific signal down a motor pathway. This type of description is reserved for the species (and even individual-specific) instantiation of that mental kind. Yet we do not want to be so general as to risk blending one mental kind with another.

Perhaps we should adopt a variant of Dennett's intentional stance. We should attempt to apply the mental kinds (described and interrelated by a psychological theory, say, folk psychology) to a given situation and 'see what sticks.' Given common-sense knowledge of mental kinds, one can proceed to attribute several mental kinds to a creature or system in a given situation. One can then judge whether each attribution helps in understanding, justifying, and predicting the behavior of the system, or does not. This method would help us relate each mental kind to environmental features and behavioral reactions that characterize them.

The problem with this method is that it smacks of behaviorism. Ironically, even as its attributes internal states, it does not tell us much about the relations between those internal states. In this sense it may run the risk of not examining the subject's psychological life closely enough, because it is not able to 'get inside' the creature. It remains at the relatively superficial level of input-output relations between it and the environment, and therefore relates mental states to one another primarily by contrasting the differing environmental features and behavioral reactions. This method goes back outside of the creature to relate (and contrast) internal states, rather than having a method of positing internal functional economies between mental states. Nevertheless it is promising because it does not seem to depend on the physical makeup of the creature or system, and in this sense may be able to identify features of the 'meta' mental kinds (those concerning regularities between environmental stimulus and behavioral reaction) which are true across species.

Other types of mental kinds – those that deal with propositional, sentential content – seem easier to attribute across species. The physical instantiation of "belief that P" should have little bearing on the content of this belief, or its relation to other sentences (i.e. its entailments). Yet even this seems an open question. Since the 'meaning,' and 'entailments,' of any proposition can only be understood in the context of its relationship to other beliefs and propositions, is seems obvious that no two individuals will hold the exact same entailments for a given proposition or belief. If this is so among individuals within the same species, surely it is even more the case for cross-species comparisons. Does the belief that P (say, the belief that everything accelerates towards the center of the earth at 9.8 meters per second squared) *mean* the same thing to each individual? Surely not, since the meaning of this statement will vary with the way (and extent to which) it is embedded with related knowledge. One can assert that two individuals believe the same thing in some limited sense, while also maintaining that this proposition means something different to them (where meaning includes the term's relation to other knowledge). Since it is probably never the case that two people share exactly the same background information, one can say two persons have the same belief just in case, *if* they had equivalent related information, they *would agree* to the same entailments. However, it seems that two subjects (given their different histories and background knowledge) can only in principle be attributed the same mental kind. For how can we say that both "Believe that P," while acknowledging that P means something different to them? We have to acknowledge that this proposition fits into their psychologies in only a partially overlapping way. Even those candidate mental kinds that seem most amenable to cross-individual and cross-species attribution – propositions and beliefs – seem almost as frustrated by individual differences as physical brain states. Yet those who offer multiple-realizability as a reason to reject type identity depend upon the possibility of mental-kind attribution across individuals, and, indeed, across species.

Human behavior belies this tentative rejection of 'meta' (cross-individual) mental kinds; after all, human beings 'successfully' attribute mental kinds to others all the time. These attributions are successful because they help us predict and understand others' behavior. Moreover, they may say that an advantage of mental kinds is that they do not demand the sort of precision that physical-level description aspires to: *of course* mental kinds will not be exactly the same in each individual; all we need is a certain degree of overlap.

This move, however, seems equally available to the reductive materialist. The materialist could say: when it comes to reduction of a mental-event token (e.g. an instance of pain at time T), the physical description should be as precise, specific, and exacting as possible. But when it comes to reduction of mental types, the physical description should be more general. The anti-reductivist will reply: Your physical description can never be general enough to include all the imaginable instantiations of a given mental kind. Physical description will fail even at its most basic elemental level if we discover a silicon-based alien, or instantiate mental kinds in a sophisticated computer.

Moreover, we need never discover the alien or build the computer to realize that abstract functional economies can be realized in *anything* of suitable complexity. Since mental kinds are not fundamentally about physical stuff, they are not similarly limited in generality.

All well and good. But the materialist will ask at this point: What is it that grounds the psychological similarities which permit attribution of the same mental kinds to two different creatures? Since our anti-reductivist purports to be a physicalist, he will have to admit that the similarities will be grounded in physical similarities: not in terms of what the creatures are made of, but in terms of how their physical stuff interacts. There must be some similarity at the dynamical level. Their psychological similarity cannot just spring out of nowhere; their physical stuff must be performing similar-enough 'operations,' we can use that word. I mean here that the physical activity of two mind-systems should be comparable and commensurable in some general way, without adverting to the specific physical quantitative and qualitative description. I do not mean that they must be instantiating some roughly comparable 'program,' since this would imply that the brain operates like a serial computer running a recursively definable program. However we do define the dynamic interaction and operations of the brain, this physical and general description need not entail the attribution of mental kinds. In short, we need not suppose that generality is limited to the basic folk-psychological thesis. How can we suppose that the potpourri of mental kinds developed by the human race over the last few millennia is the only or best way to achieve this abstract-functional-physical-dynamical description?

However, I do not take the possibility of this abstract description for granted. Why should we suppose that the same mental kind can ever be enjoyed by the same creature twice, much less by two different creatures with different histories and different physical composition? Mental activity is frightfully complex. Thus, one could reject the thesis of multiple-realizability by rejecting the possibility of type-identification at all: mental events are always unique and almost always incommensurable with any other. Therefore all we can ever have is token-identity: this specific mental event is approximate equivalent to the most accurate and complete physical description we can muster. Rejecting the possibility of type-identification, however, would be to throw the baby out with the bath water, since this would signal the impossibility of psychology as a science and discipline. Part of the reason that physical-level, reductivist description seems more grounded than folk-psychological mental description is that it cannot lie about difference. The supposed incommensurability of physical-description has long been taken to be a disadvantage, but the problems here ought to be viewed as appropriate to the difficulty of comparing any mental life across individuals.

Why Folk Psychology Appears to Work

The anti-reductivists are quick to point out the physical differences among psychological creatures, but cavalier in their assumption that mental kinds – the mental kinds of folk psychology – can be attributed across these physical differences. Understandable, given that they need this assumption to get the multiple-realizability objection going, but nevertheless objectionable in light of the foregoing discussion. Given the level of complexity here involved, our exquisite concern for physical difference seems closer to the truth about mental life than the somewhat indelicate way folk psychology glosses over these differences. Why is it so easy to be sensitive to physical difference and so predisposed to infer folk-psychological similarity and commensurability?

To reiterate a suggestion made earlier, perhaps because folk psychology is the product of tacit social agreement. We have to admit that the propensity to attribute mental kinds is a startling social-psychological fact about human beings. We reinforce and encourage one another in this attribution-activity. We tend to do it not only to other humans, but to our pets, computers, and cars, with varying levels of earnestness. Often times we have to check our tendency to anthropomorphize other animals and remind ourselves that they do not necessarily share our psychology. Owners sincerely 'explain' their misbehaving pet by claiming a dog is, 'jealous,' 'mischievous,' 'insulted,' 'hurt (emotionally),' and even 'feeling guilty.' The owner will adjust their interaction with the pet accordingly, often frustrated by the results.

Folk psychology seems to work conspicuously well in the case of normal human beings, and far less well for other systems, including abnormal human beings. Here is where the thesis of folk-psychology as tacit social agreement comes into play. Perhaps the reason why folk psychology better explains and predicts the behavior of normal human beings is because society is more adept at persuading and socializing normal examples of our species. Society and socialized individuals *make* folk psychology work precisely because they mold their behavior and mental lives to fit into its categories. This is why the human being viewing the dinner party (in Dennett's example) has an 'unfair' advantage: he is watching socialized members of his own species, who, like him, obey the same restrictions given by folk psychology.[3] He is able to reduce the myriad possibilities[4] to a relatively small set, thereby enabling him to predict the arrival of the boss and husband at the house.

Why is it is that deranged human beings do not fit so well into the folk psychological framework? Could it be because they refuse to, or cannot, listen to the story 'normal' human beings tell about themselves? Keep in mind that we do not use folk psychology only to describe and predict other's behavior. We use it to predict and explain our own. Consider that occasional feeling of surprise and dissociation one has when surprised by one's own behavior and thoughts. An unsettling feeling indeed. Now suppose every human being was confronted at an early stage of development with this feeling, even while her society was providing a ready-made story about how to understand this mysterious behavior. This behavior is *yours,* because inside you there are beliefs and desires and hopes crowding about, vying for your attention. One of those, gains the upper hand, causing you to choose the way you do.

[3] Perhaps folk psychology is Nietzsche's enemy: it makes man commensurable and predictable. The potentially boundless complexity of mental life is glossed over and restricted enough so that we become somewhat manageable for others, and manageable for ourselves. Yes, folk psychology is majestic and deep *enough*, but in settling for it we do not sound the deepest chasms. Of course, I am not advocating that folk psychology be eliminated before an alternative is available to replace it. I mean to suggest that whatever replaces FP will first have to admit that the situation is even more complex and multi-dimensional than folk psychology suggests; we may have to remain a mysterious to ourselves until later in life, until we can command this new, more powerful and sophisticated theory. Does this mean that children will use FP, or else be taught a more simplistic version of the new theory?

[4] Of course, the determinist super-physicist alien knows that it could not have been otherwise.

The human observer in Dennett's example might be at a loss if the husband, unbeknownst to him, was a schizophrenic. He might be equally, or almost, as frustrated if 'normal' (without serious chemical imbalance) human beings were not in basic agreement as to how mental life goes. On the other hand, the alien's predictive performance would not deteriorate in the slightest. The alien would give no consideration as to how the creature understands itself. This is not relevant to performing his super-physicist calculations.

Perhaps the proponents of folk psychology would prefer to say that it works within limited parameters, as do some laws or models in the hard sciences. Given certain idealizing assumptions that exclude chemically-imbalanced, immature, or demented human beings, and other species and inanimate objects, folk psychology performs well. With only these limiting assumptions, folk psychology's performance would appear impressive indeed. It would seem plausible that at some point in the future the tenets of folk psychology would become a subset of a psychological theory that includes those cases it leaves out. But there is another pivotal limiting assumption. The additional implicit assumption is that folk psychology is most accurate when applied to creatures that have been socialized into using its framework. This type of limiting assumption, I submit, is of a different character from those which exclude brain lesions or chemical imbalances. Those types of assumptions try to limit the theory to normally developed human brains. This additional assumption tries to additional limit the theory to those normally developed brains which are socialized in a certain way.[5] Is folk psychology's success impressive in such a limited domain? Should not a theory predict phenomena apart from those it was expressly engineered to explain; or ironically, predict that behavior which was molded and engineered to fit the theory? Those who cannot be so easily molded stubbornly evade folk psychology. This additional assumption gives the 'explanation' of folk psychology's success a hollow, tautological character: *of course* it works in the limited domain it was tailored for. Moreover, folk psychology's predictive and explanatory performance seems to rely upon its societal pervasiveness. If there should come a time when a new psychological framework takes its place, not only in scientific circles but in society at large, folk psychology's performance will deteriorate. But physical-level description does not similarly depend upon the vagaries and contingencies of human societies. They are valid no matter how individual human beings and societies at large understand human behavior.

That folk psychology gives consideration to how individuals understand themselves may be seen as an asset. What kind of psychology does not pay attention to how creatures understand themselves? Could such a methodology be psychology at all? The super-physicist alien, after all, does not do psychology. He only does physics. His predictions in Dennett's example are probably not seen as a special class of predictions, different from his calculations concerning inanimate objects.

But most physicalists (reductive or non-reductive) are committed to the idea that, in principle, there is an entirely adequate and complete description of any 'mental' event that is entirely physical: they are committed to token identity.

A New Model of Reduction?

[5] These comments appear to assume that there is a principled difference between 'biological development' and 'socialization,' when surely the former is in some way affected by the latter.

But ought we accept the deductive-nomological model of reduction as the second premise in this argument? The evidence of the sciences themselves seems to suggest that it would be unwise to do so. This model of reduction has never been exemplified, even in the paradigm instances of reduction witnessed in the history of science.[6] Hempel's Deductive-Nomological model of explanation long ago fell out of favor, yet the model persists in theories of reduction. (Kim, 26) We have reason to think that nomological coextensiveness is too stringent a condition for reduction. If this is so, it is no surprise that mental-to-physical reduction seems out of reach, since D-N reduction is out of reach even for agreed-upon instances of reduction in the history of science. Luckily an alternative model of reduction is in the offing.

Functional Reduction

One strategy you see often used among anti-reductivists is that they tend to speak of the 'inherent' qualities of things. They invoke the ineffable 'what its likeness' of qualia: the 'redness' of a ripe apple or the feeling of warmth when standing in the summer sun. Issues about 'qualia' are, in some ways, a separate from the status of folk psychology (whether it is a theory, whether it can be reduced, or ought to be eliminated), but I think they both issue from the same intuition. I think there is a deeply rooted belief that mental life too deep, too expansive, too varied to be 'just' the operations of a brain. With most people this leads to dualistic thinking, but in philosophical circles it leads to arguments to the effect that psychology is a sovereign, abstract science that need not make contact with the physical sciences. Putative physicalists will admit token identity while denying type-identity. Hence philosophers will agree that, *in principle*, there is an entirely accurate, totally physical description for a mental event, but they are confident that they only need to concede this in principle. After all, they will say, sentences are 'about stuff,' but the firing of neurons are not. Neurons do not have semantic content. Nor, they say, can the redness of an apple be reduced to the firing of neurons. An esteemed Emeritus Professor of Washington and Lee is fond of asking, "Where's the green?" when one declares "the brain *is* the mind." No amount of talk about color processing (e.g. what happens at the occipital lobe) seems to get us near enough to talking about the colors themselves.

The problem with all of this is that there is always a way to construe something so that it will not appear amenable to reduction. In the 19th century, the vitalists could insist that no amount of talk about metabolic processes could give a satisfactory answer to what constitutes 'life.' It took time for people to accept that life 'just was' a certain balance of metabolic processes, and not something over and above these activities. The way anti-reductivists speak of 'ineffable' qualities (like 'redness' or 'warmth') seems to be akin to the sort of strategy someone would take if they wanted to insist that what distinguishes life from death could not 'just be' certain metabolic processes. The fact is that we have

[6] Brooks cites an example given by Kitcher. "Kitcher points out that in spite of the fact that we have discovered the structure of the DNA molecule, we have in no way reduced genetics to biochemistry in the Nagelian manner…We have not succeeded in reducing laws of genetics to laws of biochemistry using bridge laws containing predicates from both sciences. This does not mean that there is something left to be explained. There are no peculiar genetic properties left dangling without a biochemical underpinning…By all intuitive standards a reduction has been accomplished." (Brooks, 804)

to 'prepare' the domain that we want to reduce for reduction; we have to start thinking about it in ways that will make the prospect of reduction seem plausible. If we persist in vitalist ways, insisting upon 'ineffable' qualities, the situation will appear hopeless.

The key, according to Kim, is to construe the property to be reduced extrinsically, rather than as some sort of 'intrinsic' quality. Suppose we want to reduce M to base properties. We must first define M in terms of the causal relationships it bears to other things. We characterize the property by virtue of its typical causes and effects; the property itself becomes a causal role. The 'base' properties – the reducers of M – will then be *occupiers* of this role. "So M is now the property of having a property with such-and-such causal potentials, and it turns out that property P is exactly the property that fits the casual specification." (Kim, 1998, p. 98) M, rather than being a distinct thing of its own, is discovered to be nothing but a second-order property of the base properties – the property of having such-and-such a causal specification. Kim uses the example of the property 'temperature.' Once we stop thinking of it as a property inhering in physical matter, and instead in terms of its causal interrelations – e.g. what happens when two bodies of different temperature are brought in contact, and what occurs to objects when they have high and low values – we are ready to find something that fits with our physical theories, that happens to satisfy these causal specifications. It turns out that 'mean molecular kinetic energy' occupies the role quite nicely. There are, then, three steps in a functional reduction. First, we must functionalize the property to be reduced. Second, we must identify the realizers of the causal role that is specified. Last, and perhaps most importantly, we must develop an explanatory theory that explains how it is that the occupiers exhibit these causal specifications.

How Does Functional Reduction Inform Study of the Mind?

Kim's proposed 'functional reduction' bears a striking resemblance to important points Paul and Patricia Churchland make about how we should proceed in theorizing about the mental. Instead of dramatizing the 'problem of consciousness' by pointing to ineffable 'what it's like' qualities of experience, the Churchlands propose to break the consciousness problem down into more manageable pieces. We need to first explicate what salient features of consciousness we would like explained, in order for the task to *even appear* as one that might bear scientific investigation. Churchland enumerates these peculiar features of consciousness in *The Engine of Reason, the Seat of the Soul*:

1. Consciousness involves short term memory.
2. Consciousness is independent of sensory inputs.
3. Consciousness displays steerable attention.
4. Consciousness has the capacity for alternative interpretations of complex/ambiguous data.
5. Consciousness disappears in deep sleep.
6. Consciousness (a confused form?) reappears in dreaming.
7. Consciousness brings together several sensory modalities into a single unified experience.

Some of these characteristics bear a striking resemblance 'causal characterization' Kim prescribes for a property we want to reduce. Even more striking is the fact that recurrent networks can exhibit all of the features mentioned in this list. Recurrent nets will feature a form of gradually-decaying short term memory, because each new incident vector is affected by vectors which have previously passed through the system. The echo effects of these vectors will dissipate as new vectors pass through. Larger recurrent networks will have more persistent short-term memories (217). Such networks can also internally generate vectors that course through the network, because recurrent projections can stimulate the 'hidden' layers sufficiently to keep the system firing. Recurrent pathways are also implicated in attention, because they can prime the system to favor the activation of certain vector prototypes: recurrent pathways can help steer the complex and noisy vector input towards a prototype as transformations on the vector are performed through the layers. Recurrent pathways can predispose the hidden layers to perform certain transformations to a portion of the vector which is incident upon them. Attention, therefore, is the disposition of the network to favor certain prototype activations. A certain predisposition is held constant against a flux of incoming data in order to rapidly pick out certain salient features. Characteristic number four – alternative interpretations – is the 'flip-side' of steerable attention. In this case, different prototype-favoring predispositions are deployed against the same set of data (vectors incident from the peripheral, sensory neurons) in order to see which achieves a better fit. (Churchland, 218-9).

Characteristics five and six are related to one another. MEG studies performed by Roldolfo Llinas show that one of the main differences between deep sleep and consciousness is the absence of large-scale, non-periodic neuronal activity, which in a waking state is tightly correlated to environmental features. This non-periodic activity is presumed to be the fingerprint of representation. In shallower, dreaming sleep, the non-periodic activity reappears, but it fails to correlate with environmental features.

None of this, of course, constitutes a knock-down explanation of these features of consciousness. Instead, it is meant to point out that for a phenomenon or problem to *look tractable to our minds*, it must be given some shape and structure. It is important to try to fix more closely what we are signifying when we say 'consciousness,' since the slippery nature of this term enables philosophers to present it as a 'quicksilver' problem that evades us as every turn. Thus suggested solutions to one feature of consciousness will be met by changing focus, and pointing to a new feature that has not yet been given 'structure.' Alternatively, they may attempt to present certain elements of consciousness as smooth-walled problems, with no possible toeholds for explanation. Thus, people will appeal to ineffable 'what-it's-like' qualities, and presuppose that that feature cannot possibly bear explanation. These philosophers will insist that the properties at issue cannot be functionalized.

The Churchlands have come out against the notion that consciousness should be presumed, at this stage of research, to constitute an especially difficult problem that is especially ill-suited to physicalist explanation. From where we stand in the current infantile state of neuroscience, it is presumptuous to expound on what problems will and will not bear out explanation. In the history of science, problems which initially appeared harder and more impenetrable than others actually fell to explanation far sooner. Guesses before 1953 about whether gene replication would be explained sooner than protein folding were just speculation, and ultimately false speculation at that. Over fifty years hence, protein folding continues to evade us. What problems look more difficult or intractable than others will always depend on the body of information we currently have at our disposal (Churchland, 1996).

The apparent plausibility of a successful reduction will depend upon how much structure the target domain is currently presumed to have. "Any reduction succeeds by reconstructing, within the resources of the new theory, the antecedently known nature, structure, and causal properties of the target phenomena. That is what inter-theoretic reduction is" (Churchland, 1996, p. 226). The problem is that those domains about which we know the least will appear to be the most resistant to reduction: "They will display no structure worth reconstructing. They will appear as a smooth-walled mystery" (226). This, however, has to do with the contingent fact of what we currently know, rather than what is possible or necessary. Moreover, Churchland grants that the existence of basic 'discriminable but structureless' representations is inevitable in any cognitive creature – right now, we call them 'qualia.' Otherwise the creature would have to discriminate representations by describing sub-features *ad infinitum*. But what representation counts as basic may well be changeable, and we have no reason to believe that what currently stand as simples will always be so. Perhaps when a new kinematics and dynamics of cognition truly takes hold, we will be able to spontaneously introspect the structure of qualia like 'redness.' We would then have new, lower-level simples. But at this juncture we have no reason to assume that 'qualia' will always remain the structureless simples of representation.

The apparently intractable nature of 'what it's like' consciousness has not been unprecedented in other scientific fields, even after a perfectly adequate explanation has been produced. The idea that some phenomenon *just is* something else entirely (with a new way of speaking germane to the reducing theory) may take time to get used to. We are prone, due to our own conceptual inertia, to believe that the domain as we have understood it has not been adequately explained by the unfamiliar terms of the new theory. "Price thus argues that phenomenal consciousness may present a case like modern physics, where it takes times and familiarity for accounts initially seen as technically adept but explanatorily unsatisfying to become accepted as genuine explanations" (Clark, 186). I am predisposed to agree with this view. Even if a completed neuroscience were in place, we might still be inclined to hold out on the 'ineffability' of qualia or consciousness. But we have to be honest about whether this is a fact about our own proclivities, or a really 'hard fact' about the ontological status of those phenomena. At any rate, supposing that we need to connect two completely distinct domains via an entirely new, basic 'psychophysical' theory, built out of the 'data' of our uncritical, introspective, subjective experience, seems a bit premature. Perhaps these 'hard problems' will fall out as neuroscience grows; but we do not know yet, and it is certainly premature to definitively rule on the possibility one way or the other.

I have been assuming that all mental kinds can be functionally characterized. For many of them, this appears to be so. But Kim suggests that those mental kinds which feature (in whole or in part) subjective qualia or conscious experience cannot be functionally construed, and therefore are not promising candidates for reduction. Kim seems to think they are not functionalizable because they are easily divorced from behavioral propensities or appropriate reaction to environmental stimuli. It seems possible that nature could have evolved creatures with no qualia or subjective experience, but who nevertheless displayed perfectly adapted tropistic behavior. In short, nature could have evolved zombies, and may actually have (perhaps insects do not have subjective experience), so why aren't all of her creatures zombies? Kim's concern about the functionalizability of qualia (and thus their amenability to reduction) seems to presuppose that a mental kind must be readily explicable in behavioral input-output terms (in relation to the environment) in order to be functionalizable. But what if qualia are 'internally functionalizable'? What if they form part of the structure of a 'black box' within the functional economy of the human being, such they input and output not to the environment, but to other subsystems within the human nervous system? Although we might grant that tropistic creatures like the sphex wasp do not require conscious experience, perhaps creatures that are as environmentally aware and functionally subtle as human beings do in fact require conscious experience in order to keep their functional economies going. Perhaps Chalmers' human zombies are not possible, precisely because qualia and subjective experience go hand-in-hand with that level of functional sophistication.

Yet many anti-reductivists maintain that consciousness does not appear to carry evolutionary advantage. For all that is important from an evolutionary perspective is that the animal have a well-calibrated functional relationship with the environment – it should be able to react to the environment in ways that will tend towards the animal's survival and propagation. But nothing about a well-calibrated functional, behavioral relationship with the environment seems to require consciousness or subjective qualia (like the redness of an apple). We could imagine that the creature could have a means for detecting apples without them having to appear so vibrantly red to its subjective consciousness. The evolutionary 'reason' for consciousness and qualia is therefore something of a mystery. Arguments about the apparent lack of evolutionary rationale for qualia and consciousness are meant to add to the general air of mystery about them. But as the foregoing suggests, I believe this argument is specious, at least in part because it presupposes the possibility of zombies. It seems more plausible to think that consciousness and qualia are a component of what happens when creatures become sophisticated enough to have memory, and achieve integrated representations of their environment.

Conclusion

Someone will not fail to point out that, over the entire course of this paper, I have been invoking the resources of folk psychology even as I have been attempting to agitate against it. True enough. Like all socially-enforced delusions, folk psychology holds such a sway over us that we can never be confident that we has begun to think outside of its categories. We have to admit that it is far more likely to be *impossible* to think without at least some of the conceptual resources it has given us. The state of our self-conception is not unlike the state of our sciences. We are at sea, trying to rebuild our ship while under still under sail. The aim of this paper has been to suggest that no matter how obvious or non-inferential certain of our beliefs may be – for instance, our belief that 'redness' is a simple and structure-less quality in our field of view, or our belief that we have beliefs (!) – we would be well-advised to admit that *every single thing* we see or think is a theory based inference. Hence every single thing is open to revision. The difficulties with folk psychology, its apparent inability to 'fit' with our broader physical theories, and the poverty of our current conceptions of how the mind works, give us ample reason to suspect that certain things we take to be fundamental are in fact ripe for revision.

Bibliography

Baumgartner, Peter and Payr, Sabine (Eds.) Speaking Minds
1995: Princeton University Press, Princeton, NJ.

Brooks, D.H.M. *How to Perform a Reduction*
Philosophy and Phenomenological Research
Vol. 54 No. 4 (Dec. 1994) pp. 803-814

Churchland, Paul M. Neurocomputational Perspective
1989: MIT Press, Cambridge, MA.

Churchland, Paul M. Matter and Consciousness
2001: MIT Press, Cambridge, MA.

Churchland, Paul M. The Engine of Reason, the Seat of the Soul:
A Philosophical Journey into the Brain
1995: MIT Press, Cambridge MA.

Churchland, Patricia S. Neurophilosophy
1986: MIT Press, Cambridge, MA.

Churchland, Patricia S. The Computational Brain
1992: MIT Press, Cambridge, MA.

Clark, Andy Mindware: An Introduction to the Philosophy of Cognitive Science
2001: Oxford UP, New York

Dennett, Daniel. “True Believers: The Intentional Strategy and Why It Works”
In Jack S. Crumley (Ed.) Problems in Mind
2000: Mayfield Publishing Company. Mountain View, CA, pp. 226-242.

Elman, Jeffrey L. “Language as a Dynamical System”
In Robert F. Port and Timothy van Gelder (Eds.) Mind as Motion
1995: Bradford Books, The MIT Press. Cambridge, MA pp. 195 – 227

Gelder, Timothy van. “Dynamics and Cognition”
In John Haugeland (Ed.) Mind Design II
1997: The MIT Press. Cambridge, MA

Gelder, Timothy van and Port, Robert F. "It's About Time: An Overview of the Dynamical Approach to Cognition" In Mind as Motion
1995: Bradford Books, The MIT Press. Cambridge, MA, pp. 1 – 45.

Jacquette, Dale. "Dualisms of Mental and Physical Phenomena"
In Jack S. Crumley (Ed.) Problems in Mind
2000: Mayfield Publishing Company. Mountain View, CA, pp. 37 – 44.

Johnson, Mark. The Body in the Mind
1987: University of Chicago Press. Chicago.

Kim, Jaegwon Mind in a Physical World
1998: The MIT Press. Cambridge, MA

Kim, Jaegwon *Supervenience, Emergence, Realization, Reduction*
Loux, Michael J. and Zimmerman, Dean W. (Eds.) The Oxford Handbook of Metaphysics 2003: Oxford UP. Oxford, UK.

Lakoff, George and Johnson, Mark. Philosophy in the Flesh
1999: Basic Books, New York.

Marras, Ausonio *Psychophysical Supervenience and Nonreductive Materialism*
1993: Synthese 93, pp. 275-304.
Netherlands: Kluwer Academic Publications

Marras, Ausonio *Kim on Reduction*
2002: Erkenntnis 57, pp. 231-257.
Netherlands: Kluwer Academic Publishers

Solso, Robert L. and Massaro, Dominic W. (Eds.) The Science of the Mind
1995: Oxford UP.